LINEAR REGRESSION

ASSIGNMENT BASED SUBJECTIVE QUESTIONS

Q1.

FROM YOUR ANALYSIS OF THE CATEGORICAL VARIABLES FROM THE DATASET, WHAT COULD YOU INFER ABOUT THEIR EFFECT ON THE DEPENDENT VARIABLE

- COUNT OF BIKE RENTALS INCREASED AND BECAME POPULAR IN YEAR 2019 THAN 2018
  (from 'Year' variable)
- COUNT OF BIKE RENTALS IS MORE DURING CLEAR WEATHER
  (from 'Weathersit' variable)
- FALL AND SUMMER ARE MORE FAVOURABLE FOR BIKE RENTALS THAN SPRING
  (from 'Season' variable)

Q2.

WHY IS IT IMPORTANT TO USE DROP_FIRST = TRUE DURING DUMMY VARIABLE CREATION?

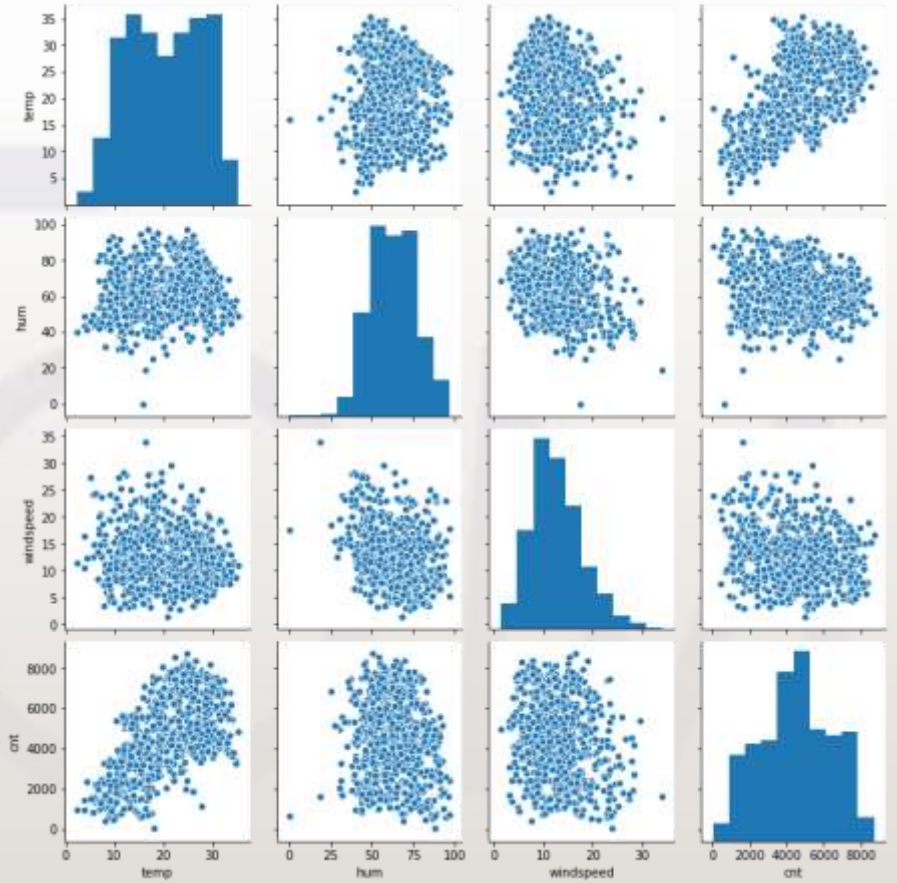
TO AVOID MULTICOLLINEARITY (IF WE DON'T DROP ,DUMMY VARIABLES WILL BE CORRELATED) AND AFFECTS THE MODEL ADVERSELY

TO AVOID REDUNDANT FEATURES

Q3.

LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?

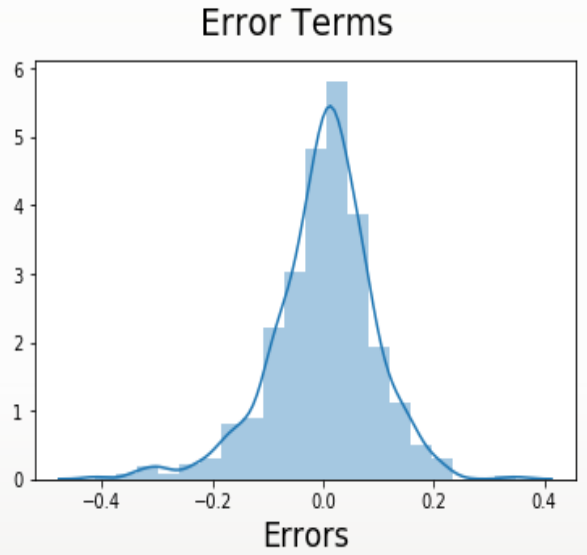
COUNT (TARGET VARIABLE)HAS SIGNIFICANTLY HIGH CORRELATION WITH TEMPERATURE (TEMP)



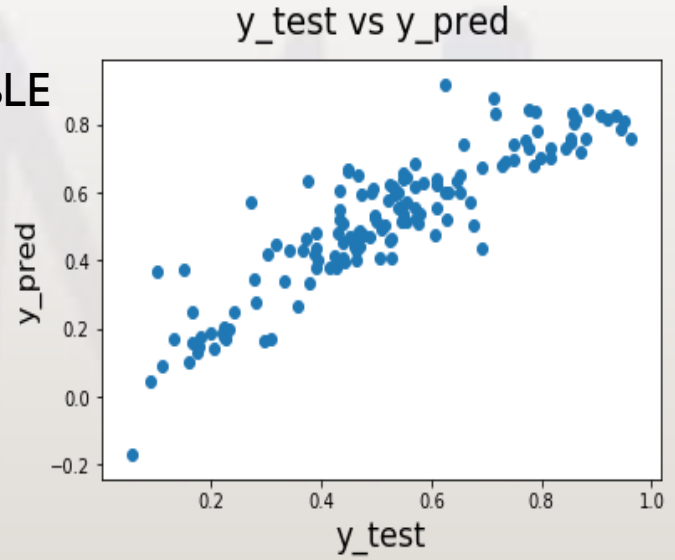
Q4.

HOW DID YOU VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET?

- RESIDUAL ERRORS FOLLOW NORMAL DISTRIBUTION



- MAINTAINS LINEAR RELATION BETWEEN DEPENDANT VARIABLE (TEST AND PREDICTED)



Q5.

BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES ?

- TEMPERATURE (0.4354)
- WEATHER SITUATION –LIGHT AND SNOWY (0.2837)
- YEAR (0.2461)

GENERAL SUBJECTIVE QUESTION

- Explain the linear regression algorithm in detail.

- Linear regression is one of the very basic forms of machine learning in the field of data science where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
- Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically ,We can write equation is:

$$y=a+bx$$

- Here, x and y are two variables on the regression line.
- b = Slope of the line.
- a = y-intercept of the line.
- x = Independent variable from dataset
- y = Dependent variable from dataset

- Explain the Anscombe's quartet in detail.

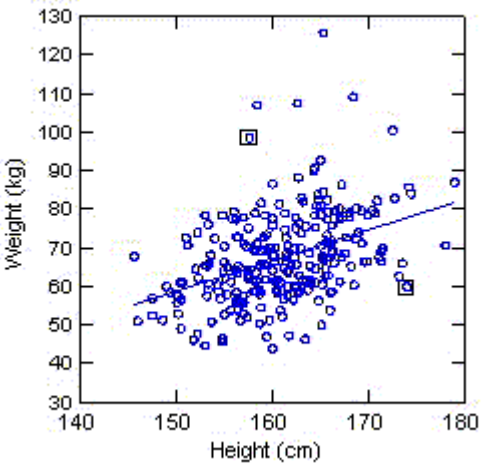
- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

- What is Pearson's R?

- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

- Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.



MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

rformed?

What is the difference between normalized scaling and standardized scaling?

- *It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.*

- *Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that* ***scaling just affects the coefficients*** *and none of the other parameters like* ***t-statistic, F-statistic, p-values, R-squared****, etc.*
- Normalization/Min max scaling: *It brings all of the data in the range of 0 and 1.* ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*

  x = x-min(x)/max(x)-min(x)

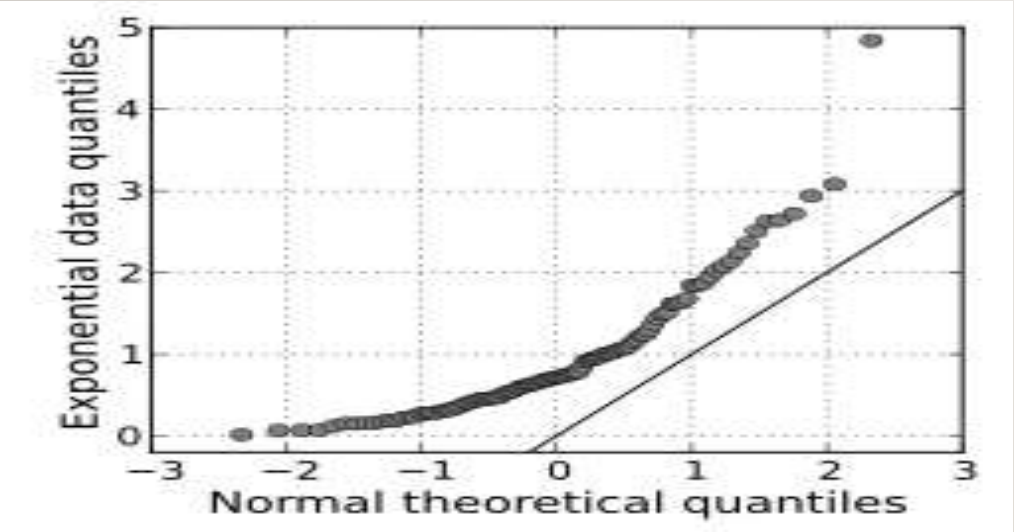
- Standardized scaling : *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

  x = x-mean(x)/sd(x)

- You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.



- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- A Q Q plot showing the 45 degree reference line:
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.